Gaussian Beta Process

by

Yingjian Wang

Department of Statistical Science
Duke University

Date: ______________________________

Approved:

____
David Dunson, Supervisor

____
Fan Li

____
Galen Reeves

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Gaussian Beta Process

by

Yingjian Wang

Department of Statistical Science
Duke University

Date: ____________________

Approved:

____
David Dunson, Supervisor

____
Fan Li

____
Galen Reeves

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2014

Abstract

This thesis presents a new framework for constituting a group of dependent completely random measures, unifying and extending methods in the literature. The dependent completely random measures are constructed based on a shared completely random measure, which is extended to the covariate space, and further differentiated by the covariate information associated with the data for which the completely random measures serve as priors. As a concrete example of the flexibility provided by the framework, a group of dependent feature learning measures are constructed based on a shared beta process, with Gaussian processes applied to build *adaptive* dependencies learnt from the practical data, denoted as the Gaussian beta process. Experiment results are presented for gene-expression series data (time as covariate), as well as digital image data (spatial location as covariate).

Contents

List of Figures

List of Abbreviations and Symbols

Abbreviations

BP	Beta process.
BPd	Beta process decomposition.
BPFA	Beta process factor analysis.
CRM	Completely random measure.
dIBP	Dependent Indian buffet process.
dHBP	Dependent hierarchical beta process.
dHDP	Dependent hierarchical Dirichlet process.
FA	Factor analysis.
GP	Gaussian process.
GBP	Gaussian beta process.
HDP	Hierarchical Dirichelet process.
HMM	Hidden Markov model.
IBP	Indian buffet process.
KBP	Kernel beta process.
KBP-FA	Kernel beta process factor analysis.
KBP-FA+	Augmented kernel beta process factor analysis.
LGM	Latent Gaussian model.
MAP	Maximum a posteriori estimation.
MCMC	Markov chain Monte Carlo.

MFCC	Mel frequency cepstral coefficients.
MGP	Multiplicative gamma process.
MSE	Mean-squared-error.
PSNR	Peak signal-to-noise ratio.
SNΓP	Spatial normalized gamma process.
TPP	Thinned Poisson process.
WGN	White Gaussian noise.

Acknowledgements

Foremost, it is my great fortune to have Prof. David Dunson as my advisor for the Master of Science degree at the Department of Statistical Science of Duke University. David is one of the greatest scientists I have ever met at Duke. His tremendous wisdom, passion, and patience in the guidance to my research are priceless treasures for me.

I would also like to thank Prof. Fan Li, whose generosity of letting me audit class STA 723 - “Case Study” is critical in helping me to improve my skill in using R. This in turn helped me a lot in the First Year Exam of the Department of Statistical Science. And thanks to Prof. Galen Reeves in giving me encouragement and helps in my research. Galen has been my committee member in both my defenses for my MS degree at DSS Dept. and PhD degree at ECE Dept. I am so happy to have Galen to witness these great moments in my life.

1

Introduction

The work presented in this thesis is proposed under the motivation to represent the possible dependencies among the real-world data in the feature learning tasks. A nonparametric tool for the feature learning is the beta process (BP), which was developed originally by Hjort (Hjort, 1990) as a Lévy process prior for "hazard measures", and was recently extended for use in feature learning (Thibaux and Jordan, 2007b), we therefore here refer to it as a "feature-learning measure." It has recently been proved (Wang and Carin, 2012) that the beta process is the limit of the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005) whose metaphor implies that the data samples serve as "customers", and the potential features serve as "dishes". The BP can also be integrated with the factor analysis model (Paisley and Carin, 2009), in which one wishes to infer a concise number of factors needed to represent the data of interest.

Based on the feature models centered with the BP, an important line of research concerns removal of the assumption of exchangeability, allowing incorporation of covariates (*e.g.*, spatial/temporal coordinates that may be available with the data). And the kernel beta process (KBP) (Ren et al., 2011a) yields an uncountable number of covariate-dependent feature-learning measures, with the beta process a special case. With the KBP, the dependencies among the real-world data are represented with the covariate-parameterized kernel functions.

In this thesis we further develop a general framework for building dependencies among a group of completely random measures (CRMs). The KBP is demonstrated to be a special case of the framework, with the corresponding Lévy measures presented. As an application of this framework, we develop a model of a group of dependent CRMs for feature learning. The model is denoted GBP, because of its use of a Gaussian process (GP) (Rasmussen and Williams, 2006) and generalization of a beta process (BP). A salutary characteristics of the GBP is that the dependencies are learned via the GP and are adapted to the data, unlike the KBP and other related kernel-based methods, for which dependencies are pre-defined by fixed kernel functions. Rather than directly imposing a kernel to modify an existing CRM, the covariates are used to place constraints on auxiliary functions, drawn in practice from a GP. These auxiliary functions are used to impose dependencies on the associated CRMs. This approach is less sensitive to selection of the kernel function (here used in the GP), as the kernel appears deeper in the model. We also develop a concise and conjugate Gibbs sampler for the inference of the GBP, by leveraging recent results from (Polson et al., 2012).

The thesis is organized as follows: first in Chapter 2 we review the definition and properties of the beta process. Next in Chapter 3 we discuss the kernel beta process designed to describe the dependencies among the data in feature learning tasks. Last in Chapter 4, we present the framework of building dependent CRMs and the GBP as a specific example. We also show the performance of the GBP on various real-world datasets, and compare with the KBP and BP.

2

Beta Process

Before the beta process is formally introduced, we briefly review the Lévy processes (Sato, 1999) and completely random measures (Kingman, 1967), which are two closely related concepts since they both demand the independence property. In fact, these two categories are overlapped with the beta process lies in their intersection.

2.1 Lévy process

2.1.1 Definition of Lévy process

A Lévy process $X(\omega)$ is a stochastic process with independent increments on a measure space $(\Omega, \mathcal{F})$. Ω is usually taken to be one-dimensional, frequently to represent a stochastic process with variation over time. A stochastic process $X(\omega)$ is a Lévy process if it satisfies the three following conditions (Applebaum, 2009):

1. $X(\varnothing) = 0$ (almost surely);
2. $X(\omega)$ has independent and stationary increments;
3. $X(\omega)$ is stochastically continuous;

In some situations we loose the second condition and also call a stochastic process with non-stationary increments as Lévy process. For the beta process example, this corresponds to the case when the concentration function $c(\omega)$ is not a constant, i.e., the inhomogeneous beta process. The beta process is reviewed in Section 2.3.

2.1.2 Pure-jump nondecreasing Lévy process and its underlying Poisson process

By the Lévy-Itô decomposition (Sato, 1999), a Lévy process can be decomposed into a continuous Brownian motion with drift, and a discrete part of a pure-jump process. When a Lévy process $X(\omega)$ only has the discrete part and its jumps are positive, then for $\forall \mathcal{A} \in \mathcal{F}$ the characteristic function of the random variable $X(\mathcal{A})$ is given by:

$$\mathbb{E}\{\mathrm{e}^{juX(\mathcal{A})}\} = \exp\{\int_{\mathbb{R}^+ \times \mathcal{A}} (\mathrm{e}^{jup} - 1)\nu(dp, d\omega)\} \tag{2.1}$$

with ν satisfying the integrability condition (Sato, 1999). The expression in (2.1) defines a category of pure-jump nondecreasing Lévy processes, including most of the Lévy processes currently used in nonparametric Bayesian methods, such as the beta, gamma, Bernoulli, and negative binomial processes. With (2.1), such a Lévy process can be regarded as a Poisson point process on the product space $\mathbb{R}^+ \times \Omega$ with the mean measure ν, called the Lévy measure. On the other hand, if the increments of $X(\omega)$ on any measurable set $\mathcal{A} \in \mathcal{F}$ are regarded as a random measure assigned on the set, then $X(\omega)$ is also a completely random measure. Due to this equivalence, in the following discussion we will not discriminate the pure-jump nondecreasing Lévy process X with its corresponding completely random measure Φ.

2.2 Completely random measure

2.2.1 Definition of completely random measure

A random measure Φ on a measure space $(\Omega, \mathcal{F})$ is termed "completely random" if for any disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_2, \cdots \in \mathcal{F}$ the random variables $\Phi(\mathcal{A}_1), \Phi(\mathcal{A}_2), \Phi(\mathcal{A}_3), \cdots$ are independent. A completely random measure Φ can be split into three independent components:

$$\Phi = \Phi_f + \Phi_d + \Phi_o \tag{2.2}$$

where $\Phi_f = \sum_{\omega \in \mathcal{I}} \phi(\omega)\delta_\omega$ is the fixed component, with the atoms in $\mathcal{I}$ fixed and the *jump* $\phi(\omega)$ random; $\mathcal{I}$ is a countable set in $\mathcal{F}$. The deterministic component Φ_d is a deterministic measure on $(\Omega, \mathcal{F})$. Φ_f and Φ_d are relatively less interesting compared to the third component Φ_o, which is called the ordinary component of Φ. According to (Kingman, 1967), Φ_o is discrete with both random atoms and jumps.

2.2.2 Lévy measure decomposition

In (Kingman, 1967), it is noted that Φ_o can be further split into a countable number of independent parts:

$$\Phi_o = \sum_k \Phi_k, \quad \Phi_k = \sum_{(\phi(\omega), \omega) \in \Pi_k} \phi(\omega)\delta_\omega \tag{2.3}$$

Denote ν as the Lévy measure of (the Lévy process corresponding to) Φ_o, ν_k as the Lévy measure of Φ_k, Π a Poisson process with ν its mean measure, and Π_k a Poisson process with ν_k its mean measure; (2.3) further yields:

$$\nu = \sum_k \nu_k, \quad \Pi = \bigcup_k \Pi_k \tag{2.4}$$

which provides a constructive method for Φ_o: first construct the Poisson process Π_k underlying Φ_k, and then with the superposition theorem (Kingman, 1993) the union of Π_k will be a realization of Φ_o.

2.3 Beta process

A beta process was first proposed by (Hjort, 1990) in survival analysis. Beta process is a Lévy process with beta-distributed increments. $B \sim \mathrm{BP}(c(\omega), \mu)$ is a beta process if

$$B(d\omega) \sim \mathrm{Beta}(c(\omega)\mu(d\omega), c(\omega)(1 - \mu(d\omega))) \tag{2.5}$$

where μ is the base measure on measure space $(\Omega, \mathcal{F})$ and a positive function $c(\omega)$ the concentration function. Expression (2.5) indicates that the increments of the beta process are independent, which makes it a special case of the Lévy process family. The Lévy measure of the beta process is

$$\nu(d\pi, d\omega) = c(\omega)\pi^{-1}(1 - \pi)^{c(\omega)-1} d\pi\mu(d\omega) \tag{2.6}$$

where $\mathrm{Beta}(0, c(\omega)) = c(\omega)\pi^{-1}(1 - \pi)^{c(\omega)-1}$ is an *improper* beta distribution since its integral over $(0, 1)$ is infinite. As a result, its *underlying Poisson process*, *i.e.*, the Poisson process with ν as its mean measure on the product space $\Omega \times (0, 1)$, denoted Π, has an infinite number of points drawn from ν, yielding

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i} \tag{2.7}$$

where π_i is the jump (increment) which happens at the atom ω_i. Real variable $\gamma = \mu(\Omega)$ is termed the mass parameter of B, and we assume $\gamma < \infty$.

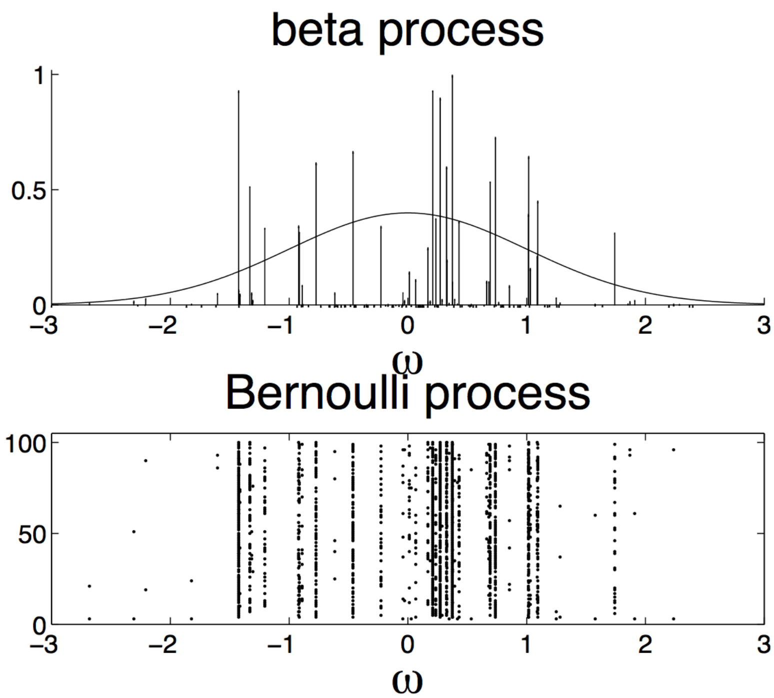

FIGURE 2.1: Beta process: **Top row:** beta process with a Gaussian base measure. **Bottom row:** 100 independent Bernoulli processes with the beta process as the prior.

3

Kernel Beta Process

KBP is a new Lévy process prior which yields an uncountable number of *covariate-dependent* feature-learning measures, with the beta process a special case (Ren et al., 2011a). This model may be interpreted as inferring covariates $\boldsymbol{x}_i^*$ for each feature (dish), indexed by i. The generative process by which the nth data sample, with covariates $\boldsymbol{x}_n$, selects features may be viewed as a two-step process. First the nth customer (data sample) decides whether to "examine" dish i by drawing $z_{ni}^{(1)} \sim \text{Bernoulli}(K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*))$, where ψ_i^* are dish-dependent kernel parameters that are also inferred (the $\{\psi_i^*\}$ defining the meaning of proximity/locality in covariate space). The kernels are designed to satisfy $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*) \in (0, 1]$, $K(\boldsymbol{x}_i^*, \boldsymbol{x}_i^*; \psi_i^*) = 1$, and $K(\boldsymbol{x}_n, \boldsymbol{x}_i^*; \psi_i^*) \to 0$ as $\|\boldsymbol{x}_n - \boldsymbol{x}_i^*\|_2 \to \infty$. In the second step, if $z_{ni}^{(1)} = 1$, customer n draws $z_{ni}^{(2)} \sim \text{Bernoulli}(\pi_i)$, and if $z_{ni}^{(2)} = 1$, the feature associated with dish i is employed by data sample n. The parameters $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i\}$ are inferred by the model. After computing the posterior distribution on model parameters, the number of kernels required to represent the measures is defined by the number of features employed from the buffet (typically small relative to the data size); this is a significant computational savings relative to (Zhou et al., 2011; Williamson et al., 2010), for which the complexity of the model is tied to the number of data samples, even if a small number of features are ultimately employed.

3.1 Review of beta-Bernoulli processes

For a beta process $B \sim \mathrm{BP}(c, B_0)$, where $c(\omega)$ is the concentration function and B_0 the base measure, the Lévy measure of $\mathrm{BP}(c, B_0)$ is given by

$$\nu(d\pi, d\omega) = c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi B_0(d\omega) \tag{3.1}$$

To draw B, one draws a set of points $(\omega_i, \pi_i) \in \Omega \times [0,1]$ from a Poisson process with measure ν, yielding

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i} \tag{3.2}$$

where δ_{ω_i} is a unit point measure at ω_i; B is therefore a discrete measure, with probability one. The infinite sum in (3.2) is a consequence of drawing Poisson(λ) atoms $\{\omega_i, \pi_i\}$, with $\lambda = \int_\Omega \int_{[0,1]} \nu(d\omega, d\pi) = \infty$. Additionally, for any set $\mathcal{A} \subset \mathcal{F}$, $B(\mathcal{A}) = \sum_{i:\ \omega_i \in \mathcal{A}} \pi_i$.

If $Z_n \sim \mathrm{BeP}(B)$ is the nth draw from a Bernoulli process, with B defined as in (3.2), then

$$Z_n = \sum_{i=1}^{\infty} b_{ni} \delta_{\omega_i}\ , \qquad b_{ni} \sim \mathrm{Bernoulli}(\pi_i) \tag{3.3}$$

A set of N such draws, $\{Z_n\}_{n=1,N}$, may be used to define whether feature $\omega_i \in \Omega$ is utilized to represent the nth data sample, where $b_{ni} = 1$ if feature ω_i is employed, and $b_{ni} = 0$ otherwise. One may marginalize out the measure B analytically, yielding conditional probabilities for the $\{Z_n\}$ that correspond to the Indian buffet process (Thibaux and Jordan, 2007b; Griffiths and Ghahramani, 2005).

3.2 Covariate-dependent Lévy process

In the above beta-Bernoulli construction, the same measure $B \sim \mathrm{BP}(c, B_0)$ is employed for generation of all $\{Z_n\}$, implying that each of the N samples have the same probabilities $\{\pi_i\}$ for use of the respective features $\{\omega_i\}$. We now assume that with each of the N samples of interest there are an associated set of covariates, denoted respectively as $\{\boldsymbol{x}_n\}$, with each $\boldsymbol{x}_n \in \mathcal{X}$. We wish to impose that if samples n and n' have similar covariates $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$, that it is probable that they will employ a similar subset of the features $\{\omega_i\}$; if the covariates are distinct it is less probable that feature sharing will be manifested.

Generalizing (3.2), consider

$$\mathcal{B} = \sum_{i=1}^{\infty} \gamma_i \delta_{\omega_i} \ , \quad \omega_i \sim B_0 \tag{3.4}$$

where $\gamma_i = \{\gamma_i(\boldsymbol{x}) : \boldsymbol{x} \in \mathcal{X}\}$ is a stochastic process (random function) from $\mathcal{X} \to [0, 1]$ (drawn independently from the $\{\omega_i\}$). Hence, $\mathcal{B}$ is a dependent *collection* of Lévy processes with the measure specific to covariate $\boldsymbol{x} \in \mathcal{X}$ being $\mathcal{B}_{\boldsymbol{x}} = \sum_{i=1}^{\infty} \gamma_i(\boldsymbol{x})\delta_{\omega_i}$. This constitutes a general specification, with several interesting special cases. For example, one might consider $\gamma_i(\boldsymbol{x}) = g\{\mu_i(\boldsymbol{x})\}$, where $g : \mathbb{R} \to [0, 1]$ is any monotone differentiable link function and $\mu_i(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ may be modeled as a Gaussian process (Rasmussen and Williams, 2006), or related kernel-based construction. To choose $g\{\mu_i(\boldsymbol{x})\}$ one can potentially use models for the predictor-dependent breaks in probit, logistic or kernel stick-breaking processes (Rodriguez and Dunson, 2009; Ren et al., 2011b; Dunson and Park, 2008). In the remainder of this chapter we propose a special case for design of $\gamma_i(\boldsymbol{x})$, termed the *kernel* beta process (KBP).

3.3 Characteristic function of the kernel beta process

Recall from Hjort (Hjort, 1990) that $B \sim \text{BP}(c(\omega), B_0)$ is a beta process on measure space $(\Omega, \mathcal{F})$ if its characteristic function satisfies

$$\mathbb{E}[e^{juB(\mathcal{A})}] = \exp\{\int_{[0,1]\times\mathcal{A}} (e^{ju\pi} - 1)\nu(d\pi, d\omega)\} \tag{3.5}$$

where here $j = \sqrt{-1}$, and $\mathcal{A}$ is any subset in $\mathcal{F}$. The beta process is a particular class of the Lévy process, with $\nu(d\pi, d\omega)$ defined as in (3.1).

For kernel $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*)$, let $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{x}^* \in \mathcal{X}$, and $\psi^* \in \Psi$; it is assumed that $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) \in [0, 1]$ for all $\boldsymbol{x}$, $\boldsymbol{x}^*$ and ψ^*. As a specific example, for the radial basis function $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) = \exp[-\psi^* \|\boldsymbol{x} - \boldsymbol{x}^*\|_2]$, where $\psi^* \in \mathbb{R}^+$. Let $\boldsymbol{x}^*$ represent random variables drawn from probability measure H, with support on $\mathcal{X}$, and ψ^* is also a random variable drawn from an appropriate probability measure Q with support over Ψ (*e.g.*, in the context of the radial basis function, ψ^* are drawn from a probability measure with support over $\mathbb{R}^+$). We now define a new Lévy measure

$$\nu_{\mathcal{X}} = H(d\boldsymbol{x}^*)Q(d\psi^*)\nu(d\pi, d\omega) \tag{3.6}$$

where $\nu(d\pi, d\omega)$ is the Lévy measure associated with the beta process, defined in (3.1).

Theorem 1 Assume parameters $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i, \omega_i\}$ are drawn from measure $\nu_{\mathcal{X}}$ in (3.6), and that the following measure is constituted

$$\mathcal{B}_{\boldsymbol{x}} = \sum_{i=1}^{\infty} \pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*)\delta_{\omega_i} \tag{3.7}$$

which may be evaluated for *any* covariate $\boldsymbol{x} \in \mathcal{X}$. For any finite set of covariates $\mathcal{S} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{|\mathcal{S}|}\}$, we define the $|\mathcal{S}|$-dimensional random vector

$\boldsymbol{K} = (K(\boldsymbol{x}_1, \boldsymbol{x}^*; \psi^*), \ldots, K(\boldsymbol{x}_{|\mathcal{S}|}, \boldsymbol{x}^*; \psi^*))^T$, with random variables $\boldsymbol{x}^*$ and ψ^* drawn from H and Q, respectively. For any set $\mathcal{A} \subset \mathcal{F}$, the $\mathcal{B}$ evaluated at covariates $\mathcal{S}$, on the set $\mathcal{A}$, yields an $|\mathcal{S}|$-dimensional random vector $\boldsymbol{\mathcal{B}}(\mathcal{A}) = (\mathcal{B}_{\boldsymbol{x}_1}(\mathcal{A}), \ldots, \mathcal{B}_{\boldsymbol{x}_{|\mathcal{S}|}}(\mathcal{A}))^T$, where $\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}) = \sum_{i:\ \omega_i \in \mathcal{A}} \pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*)$. Expression (3.7) is a covariate-dependent Lévy process with Lévy measure (3.6), and characteristic function for an arbitrary set of covariates $\mathcal{S}$ satisfying

$$\mathbb{E}[e^{j<\boldsymbol{u}, \boldsymbol{\mathcal{B}}(\mathcal{A})>}] = \exp\{\int_{\mathcal{X} \times \Psi \times [0,1] \times \mathcal{A}} (e^{j<\boldsymbol{u}, \boldsymbol{K}\pi>} - 1) \nu_{\mathcal{X}}(d\boldsymbol{x}^*, d\psi^*, d\pi, d\omega)\} \tag{3.8}$$

□

A proof is provided in the Appendix. Additionally, for notational convenience, below a draw of (3.7), valid for all covariates in $\mathcal{X}$, is denoted $\mathcal{B} \sim \text{KBP}(c, B_0, H, Q)$, with c and B_0 defining $\nu(d\pi, d\omega)$ in (3.1).

3.4 Relationship to the beta-Bernoulli process

If the covariate-dependent measure $\mathcal{B}_{\boldsymbol{x}}$ in (3.7) is employed to define covariate-dependent feature usage, then $Z_{\boldsymbol{x}} \sim \text{BeP}(\mathcal{B}_{\boldsymbol{x}})$, generalizing (3.3). Hence, given $\{\boldsymbol{x}_i^*, \psi_i^*, \pi_i\}$, the feature-usage measure is $Z_{\boldsymbol{x}} = \sum_{i=1}^{\infty} b_{\boldsymbol{x}i} \delta_{\omega_i}$, with $b_{\boldsymbol{x}i} \sim \text{Bernoulli}(\pi_i K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*))$. Note that it is equivalent in distribution to express $b_{\boldsymbol{x}i} = z_{\boldsymbol{x}i}^{(1)} z_{\boldsymbol{x}i}^{(2)}$, with $z_{\boldsymbol{x}i}^{(1)} \sim \text{Bernoulli}(K(\boldsymbol{x}, \boldsymbol{x}_i^*; \psi_i^*))$ and $z_{\boldsymbol{x}i}^{(2)} \sim \text{Bernoulli}(\pi_i)$. This model therefore yields the two-step generalization of the generative process of the beta-Bernoulli process discussed in the Introduction. The condition $z_{\boldsymbol{x}i}^{(1)} = 1$ only has a high probability when observed covariates $\boldsymbol{x}$ are near the (latent/inferred) covariates $\boldsymbol{x}_i^*$. It is deemed attractive that this intuitive generative process comes as a result of a rigorous Lévy process construction, the properties of which are summarized next.

3.5 Properties of $\mathcal{B}$

For all Borel subsets $\mathcal{A} \in \mathcal{F}$, if $\mathcal{B}$ is drawn from the KBP and for covariates $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we have

$$\begin{aligned} \mathbb{E}[\mathcal{B}_{\boldsymbol{x}}(\mathcal{A})] &= B_0(\mathcal{A})\mathbb{E}(K_{\boldsymbol{x}}) \\ \mathrm{Cov}(\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}), \mathcal{B}_{\boldsymbol{x}'}(\mathcal{A})) &= \mathbb{E}(K_{\boldsymbol{x}}K_{\boldsymbol{x}'}) \int_{\mathcal{A}} \frac{B_0(d\omega)(1 - B_0(d\omega))}{c(\omega) + 1} \\ &\quad - \mathrm{Cov}(K_{\boldsymbol{x}}, K_{\boldsymbol{x}'}) \int_{\mathcal{A}B_0^2(d\omega)} \end{aligned} \tag{3.9}$$

where, $\mathbb{E}(K_{\boldsymbol{x}}) = \int_{\mathcal{X}\times\Psi} K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*)H(dx^*)Q(d\psi^*)$. If $K(\boldsymbol{x}, \boldsymbol{x}^*; \psi^*) = 1$ for all $\boldsymbol{x} \in \mathcal{X}$, $\mathbb{E}(K_{\boldsymbol{x}}) = \mathbb{E}(K_{\boldsymbol{x}}K_{\boldsymbol{x}'}) = 1$, and $\mathrm{Cov}(K_{\boldsymbol{x}}, K_{\boldsymbol{x}'}) = 0$, and the above results reduce to the those for the original BP (Thibaux and Jordan, 2007b).

Assume $c(\omega) = c$, where $c \in \mathbb{R}^+$ is a constant, and let $\boldsymbol{K}_{\boldsymbol{x}} = (K(\boldsymbol{x}, \boldsymbol{x}_1^*; \psi_1^*), K(\boldsymbol{x}, \boldsymbol{x}_2^*; \psi_2^*), \dots)^T$ represent an infinite-dimensional vector, then for fixed kernel parameters $\{\boldsymbol{x}_i^*, \psi_i^*\}$,

$$\mathrm{Corr}(\mathcal{B}_{\boldsymbol{x}}(\mathcal{A}), \mathcal{B}_{\boldsymbol{x}'}(\mathcal{A})) = \frac{< \boldsymbol{K}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{x}'} >}{\|\boldsymbol{K}_{\boldsymbol{x}}\|_2 \cdot \|\boldsymbol{K}_{\boldsymbol{x}'}\|_2} \tag{3.10}$$

where it is assumed $< \boldsymbol{K}_{\boldsymbol{x}}, \boldsymbol{K}_{\boldsymbol{x}'} >$, $\|\boldsymbol{K}_{\boldsymbol{x}}\|_2$, $\|\boldsymbol{K}_{\boldsymbol{x}'}\|_2$ are finite; the latter condition is always met when we (in practice) truncate the number of terms used in (3.7). The expression in (3.10) clearly imposes the desired property of high correlation in $\mathcal{B}_{\boldsymbol{x}}$ and $\mathcal{B}_{\boldsymbol{x}'}$ when $\boldsymbol{x}$ and $\boldsymbol{x}'$ are proximate.

4

Gaussian beta process

For the kernel beta process discussed in Chapter 3, the underlying Poisson process of the beta process is marked with the covariate-parameterized kernel function, thus introduce dependencies represented by the covariates inherent in the data. In this chapter we show that such marking principle can be generalized than only using fixed kernel functions. To be specific, we develop a general framework for building dependencies among a group of CRMs, unifying and extending common features shared by the dependent CRMs that have emerged recently in the literature. A rigorous theoretical basis is provided for the proposed approach, with derivation of general expressions for the associated Lévy-Khinchine formula and Lévy measure.

4.1 Marked Poisson Process

In this section we briefly review the marked Poisson process. Assume there is a CRM B in the form of (2.7), with its underlying Poisson process Π of the mean measure $\nu(d\pi, d\omega)$ on the product space $\mathbb{R}^+ \times \Omega$. Then by the Marking Theorem (Kingman, 1993), with a transition probability $p(x|\pi, \omega)$, Π can be marked to form a Poisson process Π^* of the mean measure ν^* on an augmented space $\mathcal{X} \times \mathbb{R}^+ \times \Omega$:

$$\nu^*(dx, d\pi, d\omega) = p(dx|\pi, \omega)\nu(d\pi, d\omega) \tag{4.1}$$

with the points of Π^* denoted as $\{x_i, \pi_i, \omega_i\}_{i=1}^{\infty}$.

4.2 Dependent CRMs

In this section the framework of the covariate-incorporated dependent CRMs is presented. To make such incorporation possible, the underlying Poisson process of the shared CRM is extended to the covariate space with a marked Poisson process; next another marked Poisson process is applied to generate jumps (on the covariate space) differentiated by the covariates; finally a coupling function combines these jumps with the jumps of the shared CRM (on the feature space), forming the dependent CRMs.

4.2.1 Basic Framework

Covariate set: In practice we are interested in modeling N data samples, with $\{x^n\}_{n=1:N}$ the covariates associated with the data. A group of dependent CRMs $\{B^n\}_{n=1:N}$ are to be constructed to model these data samples. The locations of $\{x^n\}_{n=1:N}$ in the covariate space $\mathcal{X}$ represent the relationships between these N data samples, which imply the dependencies among $\{B^n\}_{n=1:N}$.

Two-step marked Poisson processes: Let B be the CRM and Π its underlying Poisson process shared in the construction of $\{B^n\}_{n=1:N}$, as described in Chapter 2. Since the spaces $\mathcal{X}$ and Ω are usually not same, in order to incorporate the covariate information, Π is marked to a Poisson process Π^* on the extended space $\mathcal{X} \times \mathbb{R}^+ \times \Omega$ with a transition law $p_1(x|\pi, \omega)$, as described in Section 4.1.

Then Π^* is further marked with a *covariate-parameterized* transition law, $p_2(\{g^n\}_{n=1:N}|\{x^n\}_{n=1:N}, x)$, to a Poisson process $\Pi^{\mathcal{X}}$ on the space $\mathbb{R}^N \times \mathcal{X} \times \mathbb{R}^+ \times \Omega$.

For $\Pi^{\mathcal{X}}$, its components on $\mathbb{R}^N$ are vectors $\{\boldsymbol{g}_i\}_{i=1}^{\infty}$, where $\boldsymbol{g}_i = [g_i^1, g_i^2, \cdots, g_i^N]$ are dependent jumps generated by p_2. And the point set $\{\boldsymbol{g}_i, x_i, \pi_i, \omega_i\}_{i=1}^{\infty}$ forms the Poisson process $\Pi^{\mathcal{X}}$, with the covariate information incorporated into $\boldsymbol{g}_i$. Since both p_1 and p_2 are appropriate probability laws, $\Pi^{\mathcal{X}}$ is a well-defined Poisson process by the Marking Theorem.

Coupling function: The coupling function $f : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$ is a fixed function to combined $\{\boldsymbol{g}_i\}_{i=1}^{\infty}$ and $\{\pi_i\}_{i=1}^{\infty}$ to form the dependent CRMs $\{B^n\}_{n=1:N}$ on Ω. The CRM $B^n, 1 \leqslant n \leqslant N$, can be expressed in the series form

$$B^n = \sum_{i=1}^{\infty} f(g_i^n, \pi_i)\delta_{\omega_i} \tag{4.2}$$

with $f(g_i^n, \pi_i)$ the jump of B^n at ω_i. f needs to satisfy certain conditions to guarantee $\{B^n\}_{n=1:N}$ as well-defined CRMs. Here we give a simple sufficient condition: (i) $f(\cdot, 0) = 0$ (for initial condition); (ii) non-negativity (for validity of measures); (iii) continuity (for stochastic continuity). Detailed discussion is presented in the Appendix.

Then with the covariate information of $\{x^n\}_{n=1:N}$ being integrated into the jump $f(g_i^n, \pi_i)$ of B^n, a *group* of dependent CRMs $\{B^n\}_{n=1:N}$ are constructed, where the dependencies are jointly determined by the covariates $\{x^n\}_{n=1:N}$, marked Poisson processes p_1, p_2, and coupling function f. As a concrete example, we present the dependency form of the GBP model in Section 4.3.4.

Characteristic function and Lévy measure: $\{B^n\}_{n=1:N}$ can be regarded as a Lévy process $B^{\mathcal{X}}$ with N-dimensional jumps $\boldsymbol{f}_i = [f(g_i^1, \pi_i), f(g_i^2, \pi_i), \cdots, f(g_i^N, \pi_i)]$ at ω_i, with $\Pi^{\mathcal{X}}$ its underlying Poisson process. For notation conciseness, denote $\boldsymbol{g} = [g^1, g^2, \cdots, g^N]$, $B^{\mathcal{X}}(\mathcal{A}) = [B^1(\mathcal{A}), B^2(\mathcal{A}), \cdots, B^N(\mathcal{A})]$,
$\boldsymbol{f} = [f(g^1, \pi), f(g^2, \pi), \cdots, f(g^N, \pi)]$. With the Marking Theorem, the Lévy measure

of $B^{\mathcal{X}}$ is given by

$$\nu^{\mathcal{X}} = p_2(\boldsymbol{g}|\{x^n\}_{n=1:N}, x)d\boldsymbol{g}p_1(x|\pi,\omega)dx\nu(d\pi, d\omega) \tag{4.3}$$

With the Campbell's Theorem (Kingman, 1993), the Lévy-Khinchine formula of $B^{\mathcal{X}}$ is given by

$$\begin{aligned}&\mathbb{E}[e^{j\langle \boldsymbol{u}, B^{\mathcal{X}}(\mathcal{A})\rangle}] = \\ &\exp\{\int_{\mathbb{R}^N\times\mathcal{X}\times\mathbb{R}^+\times\mathcal{A}}(e^{j\langle\boldsymbol{u},\boldsymbol{f}\rangle}-1)\nu^{\mathcal{X}}(d\boldsymbol{g}, dx, d\pi, d\omega)\}\end{aligned} \tag{4.4}$$

4.2.2 Some Examples

The framework described in Section 4.2.1 can be used to build diversified dependency forms. It unifies some existing models of dependent CRMs, such as the spatial normalized gamma process (SNΓP) (Rao and Teh, 2009), the kernel beta process (KBP) (Ren et al., 2011a), and the thinned Poisson process (Foti et al., 2013). For simplicity, we present the *marginal* Lévy measure ν^n of a B^n in the group of dependent CRMs and the coupling function f, for the KBP and thinned Poisson process.

Kernel beta process: In KBP, a fixed kernel function is applied for $\boldsymbol{g}$ which corresponds to a p_2 in the delta function form. With $\boldsymbol{g}$ marginalized out in $\mathbb{R}^N$,

$$f(g^n,\pi) = K(x^n, x, \psi)\pi \quad \nu^n = p_1(x,\psi)\nu \tag{4.5}$$

where ν is the Lévy measure of the beta process B, ψ is the precision of the kernel function K. The dependency form of KBP is fixed with $K(x^n, x_i, \psi)$. By relating atom ω_i with x_i, data with covariates x^n in the vicinity of x_i, where "vicinity" is tuned by kernel parameter ψ, are more probable to use atom ω_i.

Thinned Poisson process: The thinned Poisson process is another example of the framework when p_2 is a Bernoulli distribution,

$$\begin{aligned} f(g^n, \pi) &= g^n \pi \\ \nu^n &= [K(x^n, x)]^{g^n} [1 - K(x^n, x)]^{(1-g^n)} p_1(x|\pi, \omega)\nu \end{aligned} \tag{4.6}$$

with the probability $K(x^n, x) \in (0, 1)$. The thinned Poisson process has the nice property that the thinned CRM remains the same type of the original one, with SNΓP a special case. SNΓP also fits into the GM-dependent CRMs (Lijoi and Prünster, 2014).

4.3 Leveraging GPs

In this section we introduce the GBP model for feature learning practice, where GPs are applied on a shared BP to build adaptive dependencies.

4.3.1 GBP Model

Assume that the beta process B is drawn as represented in (2.7), from which the features ("dishes") $\{\omega_i\}$ and associated probabilities $\{\pi_i\}$ are defined. Besides we have N covariates $\{x^n\}_{n=1:N}$, which represent the N data samples ("customers"), for which the group of dependent feature learning measures $\{B^n\}_{n=1:N}$ are constructed. The features $\{\omega_i\}$ are shared across all $\{B^n\}_{n=1:N}$.

By following the framework in Section 4.2.1, for p_1 in practice we simply choose a uniform distribution on $\mathcal{X}$ as same as in (Ren et al., 2011a), since frequently the correspondence between $\{\omega_i\}$ and $\{x_i\}$ is unknown. For p_2 we choose the GP, $p_2 = \mathcal{N}(\boldsymbol{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, to elicit the adaptiveness. Here $\boldsymbol{\mu} = [m(x, x^1), m(x, x^2), \cdots, m(x, x^N)]$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariate matrix with $\Sigma_{n,n'} = k(x^n, x^{n'})$, where m and k are the mean and covariance functions of the GP. $\boldsymbol{g}$ is the N dimensional Gaussian vector evaluated at the underlying covariates $\{x^n\}_{n=1:N}$. The Lévy measure of GBP is given by

$$\nu^{\mathcal{X}} = \mathcal{N}(\boldsymbol{g}|\boldsymbol{\mu}, \boldsymbol{\Sigma})p_1(x|\pi, \omega)\nu \tag{4.7}$$

Next a sigmoid function, σ: $\mathbb{R} \mapsto [0, 1]$, is used to obtain the probability vectors $\sigma(\boldsymbol{g_i}) = [\sigma(g_i^1), \sigma(g_i^2), \cdots, \sigma(g_i^N)]$, $i = 1, 2, \cdots$. We choose σ a logistic function, leads to the coupling function of GBP,

$$f(g^n, \pi) = \frac{\pi}{1 + e^{-g^n}} \tag{4.8}$$

Obviously the f in (4.8) satisfies the conditions given in Section 4.2.1. Then $\{B^n\}_{n=1:N}$ in the form of (4.2) form a group of dependent feature learning measures. This GBP model generalizes the parametric form for a localized (in covariate space) kernel in (Ren et al., 2011a) and (Foti et al., 2013), with a nonparametric kernel form constituted via GPs to elicit adaptiveness to the data.

4.3.2 Gaussian Process on Matrix

The GPs in (4.7) can be concisely expressed by a GP on a matrix (Yu et al., 2007).

$$\begin{aligned}
&\boldsymbol{G} = [\boldsymbol{g}_1^\top, \boldsymbol{g}_2^\top, \cdots]^\top \sim \mathrm{GP}(m, [\delta, k]) \\
&\mathbb{E}(g_i^n) = m(x_i, x^n), \quad 1 \leqslant i < \infty,\ 1 \leqslant n \leqslant N \\
&\mathrm{Cov}(g_i^n, g_{i'}^{n'}) = \delta(x_i, x_{i'})k(x^n, x^{n'}), \\
&\qquad\qquad\qquad 1 \leqslant i, i' < \infty,\ 1 \leqslant n, n' \leqslant N
\end{aligned} \tag{4.9}$$

where $m : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is the mean function, $\delta, k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ are the covariance functions along the rows and columns of $\boldsymbol{G}$ respectively. $\delta(x_i, x_{i'}) = 1$, when $i = i'$; and $\delta(x_i, x_{i'}) = 0$, when $i \neq i'$, to guarantee the completely random property of $\{B^n\}_{n=1:N}$.

4.3.3 Truncation

As illustrated in (4.9), the GPs $\boldsymbol{g}_i$ are independently drawn w.r.t. $i = 1, 2, \cdots$. In practice, the number of atoms $\{\omega_i\}$ is limited to a finite number. Such truncation can be learned nonparametrically as described in the Indian buffet process (IBP) (Griffiths and Ghahramani, 2005), or calculated based on ranked jump sizes (Ferguson and Klass, 1972; Wang and Carin, 2012), or obtained by putting a threshold on the jump size (Wolpert et al., 2011). Since only a relatively small subset of $\{\pi_i\}$ are large, in practice we truncate to a large number of atoms I, and the model infers the number of atoms with significant probability π_i.

4.3.4 Adaptive Dependencies

Here we present the dependency form of GBP under conditions consistent with the model practice as shown in (4.11), where the number of features is truncated to a finite I, and the beta process B is drawn from a prior with $\pi_i \sim \text{Beta}(\alpha, \beta),\ i = 1, 2, \cdots, I$. For $\forall \mathcal{A} \in \mathcal{F}$, the correlation between B^n and $B^{n'}$ is given by

$$\text{Corr}(B^n(\mathcal{A}), B^{n'}(\mathcal{A})) = \frac{\langle \boldsymbol{\rho}^n, \boldsymbol{\rho}^{n'} \rangle}{||\boldsymbol{\rho}^n||_2 \cdot ||\boldsymbol{\rho}^{n'}||_2} \tag{4.10}$$

where the vectors $\boldsymbol{\rho}^n = [\sigma(g^n_{1*}), \sigma(g^n_{2*}), \cdots]^\top$, $\boldsymbol{\rho}^{n'} = [\sigma(g^{n'}_{1*}), \sigma(g^{n'}_{2*}), \cdots]^\top$, with atoms $\omega_{1*}, \omega_{2*}, \cdots \in \mathcal{A}$. The derivation of (4.10) and the general dependency form of GBP are given in the Appendix.

Since $\boldsymbol{\rho}^n$ and $\boldsymbol{\rho}^{n'}$ are updated by the GP in model inference, the dependencies in (4.10) are *learned* from the practical data. In fact the GP on matrix representation in (4.9) helps to reveal the relationship between the GBP and models with parametric forms of dependency, for example the KBP. In KBP, π is weighted with a fixed kernel function $K(x^n, x_i)$, which entails fixed dependencies determined by the specific form of K. While for the GBP, the GP prior is applied to elicit a *soft* kernel function adaptive to the intrinsic dependencies in the data.

4.4 Model & Inference

4.4.1 Model Description

We apply the GBP model for feature learning. Here the covariate set $\{x^n\}_{n=1:N}$ corresponds to N data samples, and $\{\omega_i\}_{i=1:I}$ are features, where I is an upper bound on the number of features used. $\{z_i^n\}_{i,n}$ is the set of binary observations, where $z_i^n = 1$ indicates that data with covariates x^n chooses feature ω_i, and $z_i^n = 0$ represents otherwise.

We construct a group of dependent CRMs $\{B^n\}_{n=1:N}$, as in Section 4.3, to represent the selection probability of customers $\{x^n\}_{n=1:N}$ upon these features $\{\omega_i\}_{i=1:I}$. π_i is the shared probability on atom ω_i, which is drawn from the BP B. $\boldsymbol{g}_i = [g_i^1, g_i^2, \cdots, g_i^N]$ decides the data-dependent tendencies of the N data samples, with covariates $\{x^n\}$, on choosing the feature ω_i. The model posterior is

$$
\begin{aligned}
p(\{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}|\{z_i^n\}) \propto & \prod_{i=1}^{I} \mathcal{N}(\boldsymbol{g}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\mathrm{Beta}(\pi_i|\alpha, \beta) \\
\times \prod_{n=1}^{N} & (\frac{\pi_i}{1+e^{-g_i^n}})^{z_i^n}(1-\frac{\pi_i}{1+e^{-g_i^n}})^{(1-z_i^n)}
\end{aligned}
\tag{4.11}
$$

where $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix for the Gaussian prior of $\boldsymbol{g}_i$, and α, β are the parameters for the beta distribution prior for π_i.

4.4.2 Model Inference

The inference of the model in (4.11) is performed via a Gibbs sampler. The difficulty in the inference comes from coupling g_i^n and π_i, which inspires two auxiliary Bernoulli random variables a_i^n and b_i^n for each observation z_i^n. To be specific,

$a_i^n \sim \text{Bernoulli}(\pi_i)$, $b_i^n \sim \text{Bernoulli}(\frac{1}{1+e^{-g_i^n}})$, which is the same idea as employed in (Ren et al., 2011a). Since $\{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ are independent w.r.t. the index i, we drop i for notational conciseness in this section.

Sample a^n and b^n: If $z^n = 1$, then $a^n = b^n = 1$. If $z^n = 0$,

$$\begin{cases} p(a^n = 0, b^n = 0 | z^n = 0) = \frac{(1-\pi)[e^{-g^n}/(1+e^{-g^n})]}{1-\pi/(1+e^{-g^n})} \\ p(a^n = 0, b^n = 1 | z^n = 0) = \frac{(1-\pi)/(1+e^{-g^n})}{1-\pi/(1+e^{-g^n})} \\ p(a^n = 1, b^n = 0 | z^n = 0) = \frac{\pi[e^{-g^n}/(1+e^{-g^n})]}{1-\pi/(1+e^{-g^n})} \end{cases}$$

Sample π: After a^n is obtained, π and $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ are decoupled. The posterior of π follows a beta-Bernoulli conjugate result: $\pi \sim \text{Beta}(\alpha + \sum_{n=1}^N a^n,\ \beta + N - \sum_{n=1}^N a^n)$.

Sample $\boldsymbol{g}$: After b^n is obtained, the part involving $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ in (4.11) is given as: $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \{b^n\}) \propto \mathcal{N}(\boldsymbol{g} | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \prod_{n=1}^N \frac{e^{-g^n(1-b^n)}}{1+e^{-g^n}}$. This conditional distribution is a latent Gaussian model (LGM) with Bernoulli-Logit likelihood. For the inference of $p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \{b^n\})$, we apply the Gibbs sampler developed in (Polson et al., 2012) which is characterized with a Pólya-Gamma distribution. First an augment variable $\boldsymbol{\lambda}$ is drawn from Pólya-Gamma distributions:

$$\lambda^n | g^n \overset{\text{ind}}{\sim} \text{PG}(1, g^n), \quad n = 1, 2, \cdots, N \tag{4.12}$$

where $\text{PG}(\cdot,\cdot)$ refers to the Pólya-Gamma distribution. Then the Gibbs sampler for updating $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ is given by:

$$\boldsymbol{\Sigma}^* = (\boldsymbol{\Lambda} + \boldsymbol{\Sigma}^{-1})^{-1}, \quad \boldsymbol{\mu}^* = \boldsymbol{\Sigma}^*(\boldsymbol{\kappa} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}) \tag{4.13}$$

where $\boldsymbol{\kappa} = [b^1 - 1/2, b^2 - 1/2, \cdots, b^N - 1/2]^T$, $\boldsymbol{\Lambda} = \text{diag}(\lambda^1, \lambda^2, \cdots, \lambda^N)$. Then $\boldsymbol{g}$ is sampled from Gaussian with the updated $\boldsymbol{\mu}^*$, $\boldsymbol{\Sigma}^*$:

$$\boldsymbol{g} | \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^* \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \tag{4.14}$$

4.5 Experiments

We perform experiments with the GBP model described in Section 4.4, considering two real-world datasets. The first one corresponds to time-dependent gene expression data extracted from human blood samples. Specifically, in a controlled study (Woods et al., 2013), humans were exposed to the H3N2 influenza virus, and the samples correspond to the time-dependent variation of the gene expression data as the host (human body) responds to the virus. In this example the covariate x^n for sample n corresponds to time from the point of virus exposure (the covariates for the different people need not be exactly at the same time, and this is accounted for via the GP).

The second class of real-world examples corresponds to analysis of patches of contiguous pixels, taken from a digital image. Here the covariate x^n is the two-dimensional spatial location of the center of patch n. In this application we consider recovery of missing color pixels ("inpainting").

For both real-world datasets, we show the adaptive dependencies that the GBP model discovers, by following (4.10), and compared with the plain BP and the KBP (Ren et al., 2011a) as a parametric way to elicit dependencies. For plain BP there is no such correlations (since covariates are not used explicitly) and for KBP the correlations are fixed due to the fixed form of the kernel function. We show the performance improvement of the GBP model over BP and KBP models for image inpainting, with the quantitative peak signal-to-noise ratio (PSNR) results presented. When presenting results, we consider different choices in the details of GBP, which underscores how it leverages covariates, and how it may be refined and tailored based on the details of the problem under study.

4.5.1 Factor Analysis For Feature Learning

The experiments on real-world data are performed with a factor analysis (FA) feature model which is reviewed here:

$$Y = D(Z \odot S) + E \tag{4.15}$$

where $Y \in \mathbb{R}^{P \times N}$ is the observed data matrix, whose columns correspond to N data samples. Matrix $D \in \mathbb{R}^{P \times I}$ is the feature matrix for a total of I features, $S \in \mathbb{R}^{I \times N}$ is the factor score matrix, $E \in \mathbb{R}^{P \times N}$ represents the noise or residual, and $\odot$ denotes the Hadamard (pointwise) vector product. We construct a FA model along the basic lines in (Zhou et al., 2009), with the priors of D, S, and E given as:

$$\begin{gathered} \omega_i \sim \mathcal{N}(0, P^{-1}\boldsymbol{I}_P) \qquad (4.16) \\ s^n \sim \mathcal{N}(0, \gamma_s^{-1}\boldsymbol{I}_I),\ \epsilon^n \sim \mathcal{N}(0, \gamma_\epsilon^{-1}\boldsymbol{I}_P) \end{gathered}$$

where ω_i, s^n, and ϵ^n are the columns of D, S, and E, $\boldsymbol{I}_P$ is the P-dimensional identity matrix, and γ_s, γ_ϵ are the precisions of s^n, ϵ^n, with diffuse $\text{Ga}(10^{-6}, 10^{-6})$ priors.

The model is slightly modified based upon the details of the data. In the context of the denoising example ϵ^n consists of a sum of Gaussian and spiky noise. In this case component j of ϵ^n is represented $\epsilon_j^n = \epsilon_{1j}^n + \epsilon_{2j}^n$, with $\epsilon_{1j}^n \sim \mathcal{N}(0, \gamma_\epsilon^{-1})$ and $\epsilon_{2j}^n \sim \zeta\delta_0 + (1-\zeta)p(\epsilon)$, for $\zeta \in (0,1)$ and $p(\epsilon)$ a uniform distribution, detailed below. This model for spiky noise was also considered in (Ren et al., 2011a).

In the BP model (Paisley and Carin, 2009), the binary matrix of dictionary usage, Z, is learned via the shared BP, *i.e.*, all N data choose the ith feature ω_i with the same probability π_i. In KBP, this probability is made data-specific, via a localized kernel function, *i.e.*, the nth data selects ω_i with probability $K(x^n, x_i, \psi)\pi_i$. We also apply the FA model described in (4.15) and (4.16) with Z learned via GBP, and make comparisons to the FA models of BP and KBP in the image-processing applications.

4.5.2 Influenza Data

We analyze H3N2 influenza gene expression data described in (Woods et al., 2013). The data are defined by $P = 12,023$ genes, sampled at $N = 14$ to $N = 16$ time instances from 17 volunteers (N is not the same for each). We set $I = 15$. The time instances (covariates) at which gene-expression data are acquired are in units of hours after inoculation, where -5 means 5 hours prior to inoculation (baseline data).

Clinical symptom scores are also given by doctors, quantified based on the severity of the symptoms developed by the subjects at a given point in time, with time points defining the covariates (see (Woods et al., 2013) for details). We wish to examine the relationship between clinical symptoms and the inferred covariate-dependent correlation, quantified via (4.10), averaged over the MCMC collection samples. Here we run 800 burn-in samples, and 200 collection samples. The 17 subjects are analyzed jointly, sharing the factor-loading matrix D, with distinct GBP prior on the feature choosing probability of each subject.

In this application of the GBP, the transition probability $p_1(x|\pi, \omega)$ in (4.7) is chosen as a uniform distribution on the covariate set, $x_i \overset{\text{ind}}{\sim} \text{Uniform}(\{x^n\}_{n=1:N})$; the prior Gaussian process in (4.11) are given as follows with c a coefficient and ψ the precision. We set $c = 0.1$, $\psi = 0.1$.

$$
\begin{aligned}
m(x_i, x^n) &= -c||x^n - x_i||_2 \\
k(x^n, x^{n'}) &= \mathrm{e}^{-\psi||x^n - x^{n'}||_2^2}
\end{aligned}
\tag{4.17}
$$

In this setup the x_i serve to locate the temporal region over which ω_i is important, with the expectation that certain factors may be important at particular (contiguous) points in time, around x_i. However, note that $m(x_i, x^n)$ is only the GP prior, and therefore the model has the flexibility to adapt the characteristics of $\boldsymbol{g}_i$, which characterizes the dependence/usage of ω_i across covariate space (time here). In fact in the experiments we observe that the GBP model is not sensitive w.r.t. the specific values of the model parameters, c and ψ. Given sufficient iterations, the GPs converge adaptively based on the data.

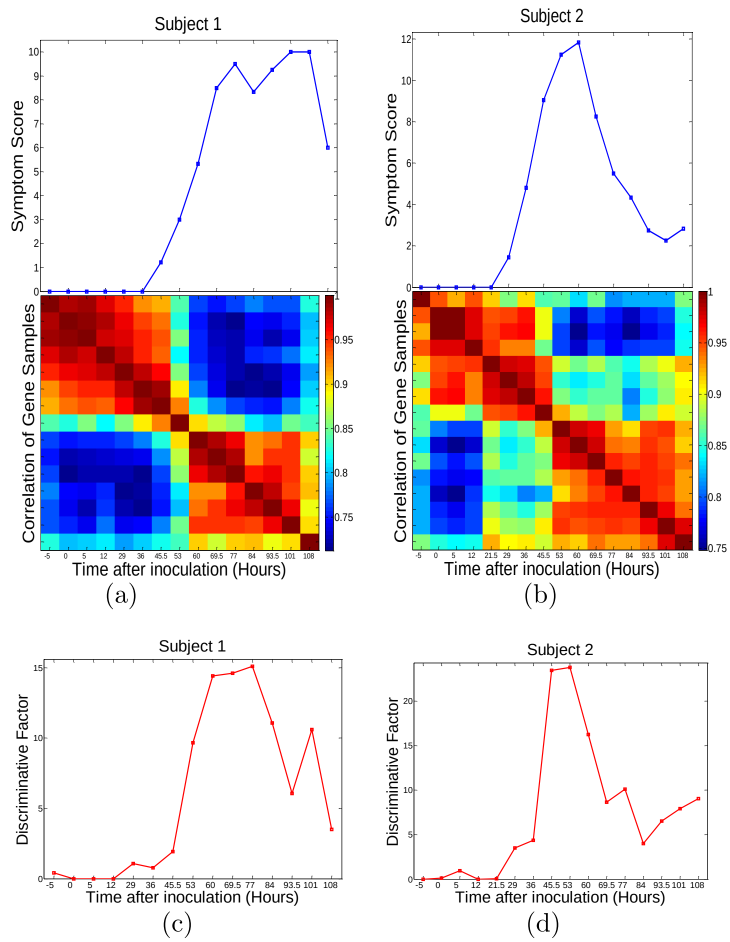

FIGURE 4.1: Results for Influenza data. (a) (b) In color at bottom, correlations of the gene expression data, at different time instances (covariates), from the posterior for two subjects. The top figures represent the associated doctor-provided symptom scores for these people. (c) (d) Discriminative factors.

For the subject considered in Figure 4.1(a), based on the inferred correlations, the 15 observed time instances may be divided into two distinct sets of times (in hours): $\{-5, 0, 5, 12, 29, 36, 45.5\}$ and $\{60, 69.5, 77, 84, 93.5, 101\}$, with the low correlations *between* the two sets indicating distinct stages in host response to the virus. This is verified by the clinical symptom scores, where the first set corresponds to the stage when there are no or limited symptoms, and the second set corresponds to the onset and persistence of symptoms.

An interesting time is 53 hours after virus exposure, around which there is a transition in the observed symptoms, and in the correlations. Another interesting time is the last point 108 hours, which shows low correlations with the second set (the person appears to be going back to the presymptom state). Similar but distinct behavior is observed for the second subject, considered in Figure 4.1(b). One of the factors, associated with a loading ω_i, appears to be particularly well linked in time to the onset of symptoms, called the "discriminative factor", shown in (c) and (d). The characteristic of the factor are consistent with the findings in (Woods et al., 2013), concerning important genes.

4.5.3 Image Inpainting

In the image processing experiments (both inpainting and denoising), we apply the GBP, KBP, and BP with the FA model described in (4.15) and (4.16). For the image processing application we wish to impose that a given atom ω_i may be used at multiple (possibly distant) regions across the image, and therefore we impose $c = 0$ to facilitate this flexibility, whereas the parametric methods like the KBP localize the usage of an ω_i in the vicinity of a particular location x_i in the sense of fixed kernels. for the covariance function of the GPs, we apply the Laplacian kernel with $\psi = 0.1$, to avoid the possible rank deficiency of the covariance matrices.

We consider a 512×512 RGB 'Barbara' image for the image inpainting demonstration. The patch size is 16×16 (resulting in $P = 768$ for the RGB data), with a total of $N = 1024$ patches. The patches are vectorized and form a 768×1024 data matrix Y. The total number of image features is set to $I = 256$.

For the image inpainting, 30% of the RGB pixels on the Barbara image are missing

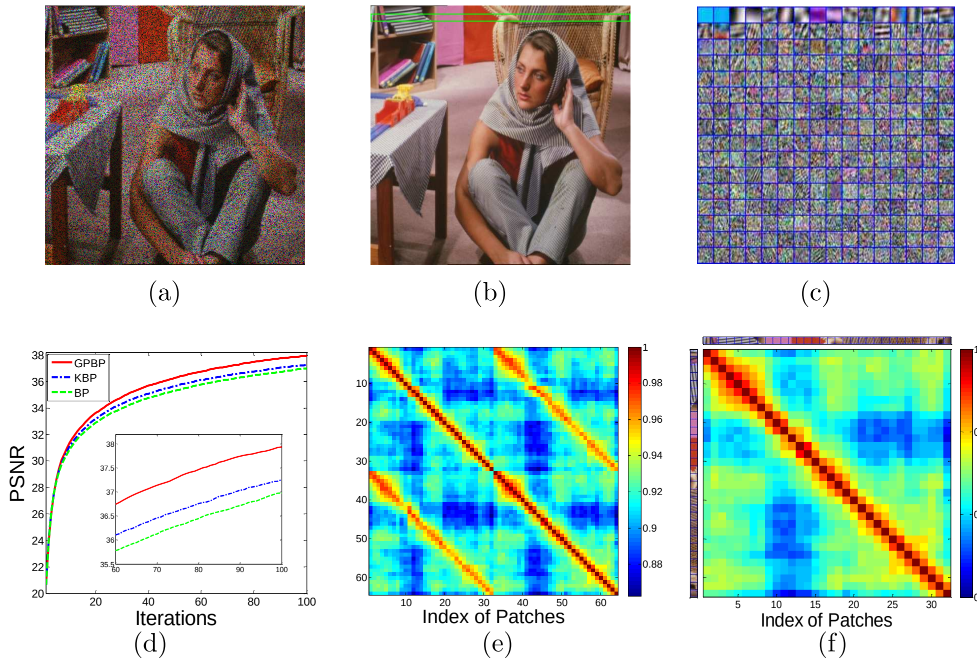

FIGURE 4.2: Image inpainting with the GBP model on the 'Barbara' image. (a) The corrupted image with 30% RGB pixels missing uniformly at random. PSNR=11.64 dB. (b) The restored image by GBP, after 100 Gibbs iterations. PSNR=37.94 dB. (c) 256 image features $\{\omega_i\}$ from the maximum-likelihood sample, ordered from top-left based on the frequency of usage. (d) Comparison of the PSNR yielded by GBP, KBP, and BP, as a function of MCMC iterations, with a zoomed-in region shown. (e) Correlation matrix of the patches in the top 2 rows of (b). (f) Correlation matrix of the patches in the second row.

uniformly at random, as shown in Figure 4.2(a), with the peak signal-to-noise ratio 11.64 (PSNR, in units of dB). (b) shows the restored image by the GBP after 100 Gibbs iteration, with PSNR 37.94. We omit the original image since the restored image essentially looks the same. Figure 4.2(c) shows the 256 image features learned via the FA model with GBP, ordered by their frequency of usage (from top-left). Note that the most widely used ω_i capture basic color and texture, while detailed but less frequently used ω_i are deeper in Figure 4.2(c) (*e.g.*, texture associated with details in the hatched scarf).

Figure 4.2(d) shows a comparison of the PSNR manifested by GBP, KBP, and BP, over the first 100 iterations. The MCMC sampler converges quickly to a good solution, from a random initialization, although clearly the number of samples is small relative to acquiring a sufficient number of samples to fully characterize the posterior (which is not our purpose). This phenomenon (fast MCMC convergence to a practically useful solution) has been observed repeatedly with such models (Zhou et al., 2009, 2011). Similar fast convergence to a practically "useful" solution was also observed for the gene data.

In comparing the performance of BP, KBP and GBP, the parameters of the FA models and model initialization are exactly the same, to guarantee fairness. In the image inpainting, the PSNR becomes stable after 500 Gibbs iterations for all the three models. We run a total of 1000 iterations and take the first 800 iterations for burn in, and average the last 200 for records, which yields the PSNR results for the GBP is 41.18, KBP is 40.87, and BP is 40.73.

The dependencies between the image patches learned via the GBP model, following (4.10), are shown in Figure 4.2(e) and (f). Again the correlations are calculated by averaging the collection samples. Subfigure (e) shows the correlations of the 2×32 image patches in the top two rows of the image. Since the indexes of the patches are labeled by row, the 33^{th} patch is right below the 1^{st} patch, resulting a high correlation between them, as clearly observed in (e).

Figure 4.2(f) shows the correlation between the 32 image patches in the second row, which is labeled as in (b). We put the 32 patches along the two sides of the correlation matrix to make a clear observation. In (f), there are two conspicuously high-correlated groups observed. The first group includes the patches from the books on the shelf, corresponding to the upper-left part on the correlation matrix. The high correlations come from the similarities between these patches. The second group comes from the patches of the chair, corresponding to the lower-right part on the correlation matrix. There are also strong correlations between the pure pink and red patches.

4.6 Summary

A unifying framework for building dependent CRMs by leveraging covariate information is presented. Different with the hierarchical constructions as in HDP (Teh et al., 2006) and HBP (Thibaux and Jordan, 2007a) where the sharing of atoms is accomplished by inheriting from parent processes, this framework applies various marked Poisson processes parameterized by covariates. Not confined to multiplication form in (4.8), diversified coupling functions, e.g. exponential forms of π, can be used to facilitate different types of dependencies constructed within the framework.

As a concrete example the GBP model is developed for adaptively learning dependencies in the data, enriching the usage of GP. The inference of the GBP is performed with a fully analytic Gibbs sampler, with adaptive dependencies discovered and superior quantitative results yielded. The GBP model may also be extended to *non-completely* random cases by replacing the delta kernel function in (4.9) and thus introducing dependencies between the features. This extension can be especially useful for modeling problems where the data are better described by stochastic processes with partially dependent increments.

Appendix A

Gaussian beta process

A.1 The framework of dependent CRMs

A.1.1 Background

Lévy process

A Lévy process is essentially a stochastic process with independent increments (Applebaum, 2009). Let X be a stochastic process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For simplicity assume $\Omega = \mathbb{R}^+$ which may correspond to time, and $\mathcal{F}$ is the Borel field on $\mathbb{R}^+$. We say X is a Lévy process if:

(L1) $X(0) = 0$ almost surely;

(L2) For any $0 \leqslant t_1 \leqslant \cdots \leqslant t_N \leqslant \infty$, $X(t_n - t_{n-1})$ and $X(t_{n-1} - t_{n-2})$, $2 \leqslant n \leqslant N$, are independent;

(L3) X is stochastically continuous, i.e., for $\forall c > 0$ and $\forall s \geqslant 0$

$$\lim_{t \to s} \mathbb{P}(|X(t) - X(s)| > c) = 0 \tag{A.1}$$

The characteristic function of a Lévy process X is given by the Lévy-Khintchine formula: for $\forall t \geqslant 0$,

$$\mathbb{E}(e^{juX(t)}) = \exp\{jaut - \frac{1}{2}\sigma^2 tu^2 + t\int_{\mathbb{R}\backslash\{0\}}[e^{jux} - 1 - jux\mathrm{I}(|x| < 1)]\nu(dx)\} \quad \text{(A.2)}$$

where $a \in \mathbb{R}$, $\sigma \geqslant 0$, $\mathrm{I}(\cdot)$ is the indicator function, and ν is the Lévy measure which satisfies the condition

$$\int_{\mathbb{R}\backslash\{0\}} \min(x^2, 1)\nu(dx) < \infty \quad \text{(A.3)}$$

A Lévy process X can be decomposed into three independent components, a linear drift, a Brownian motion and a compound Poisson process by the Lévy-Itō decomposition, which results the Lévy-Khintchine triplet $\{a, \sigma^2, \nu\}$. Note that in the three laws of the Lévy process the stationary condition is not required. So here the term 'Lévy process' also includes the inhomogeneous Lévy process, such as the beta process whose concentration is not constant.

Completely random measure

A set function μ on a measurable space $(\Omega, \mathcal{F})$ is a *measure* if it satisfies the three conditions below (Billingsley, 1995):
(M1) For $\forall \mathcal{A} \in \mathcal{F}$, $\mu(\mathcal{A}) \in [0, \infty]$;
(M2) $\mu(\varnothing) = 0$;
(M3) if $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots$ are disjoint measurable sets in $\mathcal{F}$, and if $\cup_{n=1}^{\infty}\mathcal{A}_n \in \mathcal{F}$, then $\mu(\cup_{n=1}^{\infty}\mathcal{A}_n) = \sum_{n=1}^{\infty}\mu(\mathcal{A}_n)$.

A random measure Φ on a measure space $(\Omega, \mathcal{F})$ is termed 'completely random' if for any disjoint sets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \cdots \in \mathcal{F}$ the random variables $\Phi(\mathcal{A}_1), \Phi(\mathcal{A}_2), \Phi(\mathcal{A}_3), \cdots$ are independent (Kingman, 1967). A completely random measure (CRM) Φ can be split into three independent components:

$$\Phi = \Phi_f + \Phi_d + \Phi_o \quad \text{(A.4)}$$

where $\Phi_f = \sum_{\omega \in \mathcal{I}} \phi(\omega)\delta_\omega$ is the fixed component, with the atoms in $\mathcal{I}$ fixed and the *jump* $\phi(\omega)$ random; $\mathcal{I}$ is a countable set in $\mathcal{F}$. The deterministic component Φ_d is a deterministic measure on $(\Omega, \mathcal{F})$. Φ_f and Φ_d are relatively less interesting compared to the third component Φ_o, which is called the ordinary component of Φ. According to (Kingman, 1967), Φ_o is discrete with both random atoms and jumps.

The category of the pure-jump non-decreasing Lévy processes also fits into the family of CRM (Kingman, 1967) (Kingman, 1993). The beta process, gamma process, Bernoulli process all belong to this category. The Lévy processes/CRMs in this category have simple forms of the Lévy-Khintchine formula.

We use the beta process as an example. Let $B \sim \mathrm{BP}(c(\omega), \mu(d\omega))$ be a beta process on $(\Omega, \mathcal{F})$. Since B only has jump increments, the drift term and Brownian motion term in its Lévy-Khintchine representation disappear. And since B is non-decreasing, i.e. B is a subordinator, the term $jux\mathrm{I}(|x| < 1)$ in the integral of (A.2) is also gone. Then the Lévy-Khintchine formula of the beta process B is given by, for $\forall \mathcal{A} \in \mathcal{F}$,

$$\begin{aligned} \mathbb{E}[e^{juB(\mathcal{A})}] &= \exp\{\int_{[0,1]\times\mathcal{A}} (e^{ju\pi} - 1)\nu(d\pi, d\omega)\} \\ \nu(d\pi, d\omega) &= c(\omega)\pi^{-1}(1-\pi)^{c(\omega)-1} d\pi\mu(d\omega) \end{aligned} \tag{A.5}$$

where ν is the Lévy measure of the beta process B. Note the Lévy measure ν in (A.5) also comprises the base measure $\mu(d\omega)$ on Ω, which is different with the Lévy measure shown in (A.2) where the Lesbegue measure on $\mathbb{R}^+$ is implicitly applied as the base measure.

The Lévy measure ν in (A.5) can be regarded as the mean measure of the *underlying Poisson process* of B, denoted as Π. Since ν is characterized with an *improper* beta distribution, whose integral over $(0, 1)$ is infinite, Π has an infinite number of points drawn from ν, denoted as $\{\pi_i, \omega_i\}_{i=1}^{\infty}$, yielding the series expression of the beta process:

$$B = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i} \tag{A.6}$$

where π_i is the jump which happens at atom ω_i.

Marked Poisson process

Assume there is a CRM B in the form of (A.6), with its underlying Poisson process Π of the mean measure $\nu(d\pi, d\omega)$ on the product space $\mathbb{R}^+ \times \Omega$. Then by the Marking Theorem (Kingman, 1993), with a transition probability $p(x|\pi, \omega)$, Π can be marked to form a Poisson process Π^* of the mean measure ν^* on an augmented space $\mathcal{X} \times \mathbb{R}^+ \times \Omega$:

$$\nu^*(dx, d\pi, d\omega) = p(dx|\pi, \omega)\nu(d\pi, d\omega) \tag{A.7}$$

with the points of Π^* denoted as $\{x_i, \pi_i, \omega_i\}_{i=1}^{\infty}$.

A.1.2 Basic framework

Our goal is to build a group of dependent CRMs $\{B^n\}_{n=1:N}$ on Ω, based on a *shared* CRM B, and a covariate set $\{x^n\}_{n=1:N}$ which naturally comes up with the real-world data.

Covariate set

Assume we have a set of *fixed* covariates in the covariate space $\mathcal{X}$, denoted as $\{x^n\}_{n=1:N}$. The covariates typically correspond to the spatial information borne in the real-world data, e.g. the time instances of the gene expression ($\mathcal{X} = \mathbb{R}^+$), locations of the patches on an image plane ($\mathcal{X} = \mathbb{R}^2$), etc. The N covariates $\{x^n\}_{n=1:N}$ are associated with the N data samples, for which we try to build N dependent CRMs on Ω, $\{B^n\}_{n=1:N}$, whose dependencies are reflected by the relationships between the covariates $\{x^n\}_{n=1:N}$.

Two-step marked Poisson processes

Since the covariate space $\mathcal{X}$ and the Ω are usually not the same, the underlying Poisson process Π of B is marked to a Poisson process Π^* on $\mathcal{X} \times \mathbb{R}^+ \times \Omega$ by a transition law $p_1(x|\pi, \omega)$, as described in Section A.1.1, to allow the incorporation of the covariate information, through the 'bridge' $\{x_i\}_{i=1}^{\infty}$.

To achieve this incorporation, Π^* is further marked with a *covariate-parameterized* transition probability, $p_2(\{g^n\}_{n=1:N}|\{x^n\}_{n=1:N}, x)$, to a Poisson process $\Pi^{\mathcal{X}}$ on the product space $\mathbb{R}^N \times \mathcal{X} \times \mathbb{R}^+ \times \Omega$. Since both the transition p_1 and p_2 are appropriate probability laws, $\Pi^{\mathcal{X}}$ is a well-defined Poisson process by the Marking theorem (Kingman, 1993).

For $\Pi^{\mathcal{X}}$, its components on $\mathbb{R}^N$ are vectors $\{\boldsymbol{g}_i\}_{i=1}^{\infty}$, where $\boldsymbol{g}_i = [g_i^1, g_i^2, \cdots, g_i^N]$. And the point set $\{\boldsymbol{g}_i, x_i, \pi_i, \omega_i\}_{i=1}^{\infty}$ forms the Poisson process $\Pi^{\mathcal{X}}$, with the covariate information incorporated into $\boldsymbol{g}_i$. There are correlations between the elements of $\boldsymbol{g}_i$, $g_i^1, g_i^2, \cdots, g_i^N$, which are introduced by the transition probability p_2. And such correlations are combined with the jump of B, π_i, by the coupling function to form the dependent CRMs $\{B^n\}_{n=1:N}$.

Coupling function

The coupling function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is a fixed function to combined $\{\boldsymbol{g}_i\}_{i=1}^{\infty}$ and $\{\pi_i\}_{i=1}^{\infty}$ to form the dependent CRMs $\{B^n\}_{n=1:N}$. The CRM $B^n, 1 \leqslant n \leqslant N$, can be expressed as

$$B^n = \sum_{i=1}^{\infty} f(g_i^n, \pi_i)\delta_{\omega_i} \tag{A.8}$$

with $f(g_i^n, \pi_i)$ the jump of B^n at ω_i. $\{B^n\}_{n=1:N}$ can be regarded as a Lévy process with N-dimensional jumps $\boldsymbol{f}_i = [f(g_i^1, \pi_i), f(g_i^2, \pi_i), \cdots, f(g_i^N, \pi_i)]$ at ω_i.

The coupling function f needs to meet certain conditions to make $\{B^n\}_{n=1:N}$

well-defined Lévy processes. Here we give a set of sufficient conditions of f for the validity of $\{B^n\}_{n=1:N}$ as Lévy processes, with recalling the three laws of the Lévy process given in Section A.1.1. For simplicity of discussion, note that $G^n = \sum_{i=1}^{\infty} g_i^n \delta_{x_i}$ is also a Lévy process, here named as the 'auxiliary process'.

For (L1), we need $f(g_i^n, \pi_i) = 0$ when $\pi_i = 0$; For (L2), since both $\{g_i^n\}_{i=1}^{\infty}$ and $\{\pi_i\}_{i=1}^{\infty}$ are independent w.r.t. the index i, $\{f(g_i^n, \pi_i)\}_{i=1}^{\infty}$ are also independent w.r.t. i; For (L3), since both G^n and B are Lévy processes with stochastically continuous increments, then by the Continuous mapping theorem (Billingsley, 1968), as long as f is continuous, B^n is also stochastically continuous. Note since continuity is true, B^n are almost surely finite on any bounded set. The sufficient conditions of f for $\{B^n\}_{n=1:N}$ to be well-defined Lévy processes is given as below:

(A1) $f(\cdot, 0) = 0$;

(A2) $f(\cdot, \cdot)$ is continous on $\mathbb{R} \times \mathbb{R}$;

For the validity of $\{B^n\}_{n=1:N}$ as CRMs, we recall the three laws of CRMs given in Section A.1.1. For (M1), $f(g_i^n, \pi_i)$ needs to be nonnegative on $\mathbb{R} \times \mathbb{R}$, and almost surely finite on any bounded set; For (M2), there is $f(g_i^n, \pi_i) = 0$ when $\pi_i = 0$, same as the condition (A1); For (M3), since B is pure-jump and f is applied on the jumps of $\{B^n\}_{n=1:N}$, (M3) is naturally satisfied. The sufficient conditions of f for $\{B^n\}_{n=1:N}$ to be well-defined CRMs is given as below:

(B1) $f(\cdot, 0) = 0$;

(B2) $f(\cdot, \cdot)$ is non-negative on $\mathbb{R} \times \mathbb{R}$;

Then combine the conditions in (A) and (B), plus the fact that continuity guarantees almost surely finiteness on any bounded set, a sufficient set of conditions of

f for $\{B^n\}_{n=1:N}$ to be both well-defined Lévy processes as well as CRMs is given as below:

(C1) $f(\cdot, 0) = 0$;

(C2) $f(\cdot, \cdot)$ is non-negative on $\mathbb{R} \times \mathbb{R}$;

(C3) $f(\cdot, \cdot)$ is continous on $\mathbb{R} \times \mathbb{R}$;

Characteristic function and Lévy measure of $\{B^n\}_{i=1:N}$

If we regard the N dependent CRMs $\{B^n\}_{n=1:N}$ as a Lévy process with N-dimensional jumps, denoted as $B^{\mathcal{X}}$, then the Lévy measure of $B^{\mathcal{X}}$, denoted as $\nu^{\mathcal{X}}$, is also the mean measure of its underlying Poisson process $\Pi^{\mathcal{X}}$. With the Marking theorem, $\nu^{\mathcal{X}}$ is given by

$$\nu^{\mathcal{X}}(d\boldsymbol{g}, dx, d\pi, d\omega) = p_2(\boldsymbol{g}|\{x^n\}_{n=1:N}, x)d\boldsymbol{g}p_1(x|\pi, \omega)dx\nu(d\pi, d\omega) \tag{A.9}$$

For notation conciseness, denote $\boldsymbol{g} = [g^1, g^2, \cdots, g^N]$,
$B^{\mathcal{X}}(\mathcal{A}) = [B^1(\mathcal{A}), B^2(\mathcal{A}), \cdots, B^N(\mathcal{A})]$, $\boldsymbol{f} = [f(g^1, \pi), f(g^2, \pi), \cdots, f(g^N, \pi)]$. With the Campbell's theorem (Kingman, 1993), the Lévy-Khintchine formula of $B^{\mathcal{X}}$ is given by

$$\mathbb{E}[e^{j\langle \boldsymbol{u}, B^{\mathcal{X}}(\mathcal{A})\rangle}] = \exp\{\int_{\mathbb{R}^N \times \mathcal{X} \times \mathbb{R} \times \mathcal{A}} (e^{j\langle \boldsymbol{u}, \boldsymbol{f}\rangle} - 1)\nu^{\mathcal{X}}(d\boldsymbol{g}, dx, d\pi, d\omega)\} \tag{A.10}$$

The dependence between B^n and $B^{n'}$, $1 \leqslant n, n' \leqslant N$, is jointly determined by the covariates $\{x^n\}_{n=1:N}$, marked Poisson processes p_1, p_2, and the coupling function f. As a specific example of the framework of dependent CRMs presented in this paper, we show the Lévy-Khintchine formula and the correlation form for the GPBP.

A.2 Correlation between B^n and $B^{n'}$

Given a beta process $B \sim \mathrm{BP}(c(\omega), \mu(d\omega))$, and the $B^n, B^{n'}$ following the GPBP construction presented in main manuscript, $\forall\ \mathcal{A} \in \mathcal{F}$, the correlation between $B^n(\mathcal{A})$ and $B^{n'}(\mathcal{A})$ is given by

$$\begin{aligned}
\mathrm{Corr}(B^n(\mathcal{A}), B^{n'}(\mathcal{A})) &= \frac{\int_{\mathcal{A}} \mathrm{Cov}(\rho^n(x)B(d\omega), \rho^{n'}(x)B(d\omega))}{\{\int_{\mathcal{A}} \mathrm{Var}(\rho^n(x)B(d\omega)) \cdot \int_{\mathcal{A}} \mathrm{Var}(\rho^{n'}(x)B(d\omega))\}^{\frac{1}{2}}} \\
&= \frac{\sum_{i:\omega_i \in \mathcal{A}} \rho_i^n \rho_i^{n'} \mathrm{Var}(\pi_i)}{\{\sum_{i:\omega_i \in \mathcal{A}} (\rho_i^n)^2 \mathrm{Var}(\pi_i) \cdot \sum_{i:\omega_i \in \mathcal{A}} (\rho_i^{n'})^2 \mathrm{Var}(\pi_i)\}^{\frac{1}{2}}}
\end{aligned} \tag{A.11}$$

In (A.11) we observe that $\boldsymbol{\rho^n}, \boldsymbol{\rho^{n'}}$, where the vectors $\boldsymbol{\rho}^n = [\sigma(g_{1*}^n), \sigma(g_{2*}^n), \cdots]^\top$, $\boldsymbol{\rho}^{n'} = [\sigma(g_{1*}^{n'}), \sigma(g_{2*}^{n'}), \cdots]^\top$, with atoms $\omega_{1*}, \omega_{2*}, \cdots \in \mathcal{A}$ and $1^*, 2^*, \cdots$ are the index of atoms falling in $\mathcal{A}$, play the role to tune the correlation between B^n and $B^{n'}$, by the adaptiveness of the Gaussian processes. In practice as described in the GPBP model construction in the main manuscript, we assume that the number of features is truncated to a finite I, and the beta process B is drawn from a prior with $\pi_i \sim \mathrm{Beta}(\alpha, \beta),\ i = 1, 2, \cdots, I$. In this case, the correlation between B^n and $B^{n'}$ can be simplified as the cosine of the angle between the vectors $\boldsymbol{\rho^n}, \boldsymbol{\rho^{n'}}$.

$$\begin{aligned}
\mathrm{Corr}(B^n(\mathcal{A}), B^{n'}(\mathcal{A})) &= \frac{\mathrm{Var}(\pi_i) \sum_{i:\omega_i \in \mathcal{A}} \rho_i^n \rho_i^{n'}}{\mathrm{Var}(\pi_i)\{\sum_{i:\omega_i \in \mathcal{A}} (\rho_i^n)^2 \cdot \sum_{i:\omega_i \in \mathcal{A}} (\rho_i^{n'})^2\}^{\frac{1}{2}}} \\
&= \frac{\sum_{i:\omega_i \in \mathcal{A}} \rho_i^n \rho_i^{n'}}{\{\sum_{i:\omega_i \in \mathcal{A}} (\rho_i^n)^2 \cdot \sum_{i:\omega_i \in \mathcal{A}} (\rho_i^{n'})^2\}^{\frac{1}{2}}} \\
&= \frac{< \boldsymbol{\rho^n}, \boldsymbol{\rho^{n'}} >}{\| \boldsymbol{\rho^n} \|_2 \cdot \| \boldsymbol{\rho^{n'}} \|_2}
\end{aligned} \tag{A.12}$$

Bibliography

Applebaum, D. (2009), *Levy Processes and Stochastic Calculus*, Cambridge University Press.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

Billingsley, P. (1995), *Probability and Measure*, Wiley-Interscience, 3 edn.

Dunson, D. B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307–323.

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Ferguson, T. and Klass, M. (1972), "A Representation of Independent Increment Processes without Gaussian Components," *The Annals of Mathematical Statistics*.

Foti, N., Futoma, J., Rockmore, D., and Williamson, S. (2013), "A unifying representation for a class of dependent random measures," in *AISTATS*.

Griffiths, T. and Ghahramani, Z. (2005), "Infinite latent feature models and the Indian buffet process," in *NIPS*.

Hjort, N. (1990), "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data," *Annals of Statistics*.

Jordan, M. (2009), "Hierarchical models, nested models and completely random measures," in *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, New York: Springer.

Kingman, J. (1967), "Completely Random Measure," in *Pacific Journal of Mathematics*, vol. 21(1):59-78.

Kingman, J. (1993), *Poisson Processes*, Oxford University Press, Oxford.

Knowles, D. and Ghahramani, Z. (2007), "Infinite Sparse Factor Analysis and Infinite Independent Components Analysis," in *Independent Component Analysis and Signal Separation*.

Lijoi, A. and Prünster, I. (2014), "Bayesian inference with dependent normalized completely random measures," *Bernoulli*.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *In Proceedings of the Section on Bayesian Statistical Science*.

Miller, K., Griffiths, T., and Jordan., M. I. (2008), "The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features," in *UAI*.

Paisley, J. and Carin, L. (2009), "Nonparametric Factor Analysis with Beta Process Priors," in *ICML*.

Paisley, J., Zaas, K., Woods, C., Ginsburg, G., and Carin, L. (2010), "A Stick-Breaking Construction of the Beta Process." in *ICML*, pp. 847–854.

Polson, N., Scott, J., and Windle, J. (2012), "Bayesian inference for logistic models using Polya-Gamma latent variables, http://arxiv.org/abs/1205.0310," .

Rao, V. and Teh, Y. (2009), "Spatial Normalized Gamma Processes," in *NIPS*.

Rasmussen, C. and Williams, C. (2006), *Gaussian Processes for Machine Learning*, MIT Press.

Ren, L., Dunson, D., Lindroth, S., and Carin, L. (2010), "Dynamic Nonparametric Bayesian Models for Analysis of Music," *Journal of The American Statistical Association*, 105, 458–472.

Ren, L., Wang, Y., Dunson, D., and Carin, L. (2011a), "The Kernel Beta Process," in *NIPS*.

Ren, L., Du, L., Carin, L., and Dunson, D. B. (2011b), "Logistic Stick-Breaking Process," *J. Machine Learning Research*.

Rodriguez, A. and Dunson, D. B. (2009), "Nonparametric Bayesian models through probit stickbreaking processes," *Univ. California Santa Cruz Technical Report*.

Sato, K. (1999), *Lévy processes and infinitely divisible distributions*, Cambridge University Press.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006), "Hierarchical Dirichlet processes," *JASA*.

Thibaux, R. and Jordan, M. (2007a), "Hierarchical beta processes and the Indian buffet process," in *AISTATS*.

Thibaux, R. and Jordan, M. I. (2007b), “Hierarchical beta processes and the Indian buffet process,” in *AISTATS*.

Wang, Y. and Carin, L. (2012), “Levy Measure Decompositions for the Beta and Gamma Processes,” in *ICML*.

Williamson, S., Orbanz, P., and Ghahramani, Z. (2010), “Dependent Indian buffet processes,” in *AISTATS*.

Wolpert, R., Clyde, M., and Tu, C. (2011), “Stochastic Expansions using Continuous Dictionaries: Lévy Adaptive Regression Kernels,” *Annals of Statistics*.

Woods, C., McClain, M., Chen, M., Zaas, A., Nicholson, B., Varkey, J., Veldman, T., Kingsmore, S., Huang, Y., Lambkin-Williams, R., Gilbert, A., Ramsburg, A. H. E., Glickman, S., Lucas, J., Carin, L., and Ginsburg, G. (2013), “A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2,” *PLoS ONE*.

Yu, K., Chu, W., Yu, S., Tresp, V., and Xu, Z. (2007), “Stochastic relational models for discriminative link prediction,” in *NIPS*.

Zhou, M., Chen, H., Paisley, J., Ren, L., Sapiro, G., and Carin, L. (2009), “Non-Parametric Bayesian Dictionary Learning for Sparse Image Representations,” in *NIPS*.

Zhou, M., Yang, H., Sapiro, G., Dunson, D., and Carin, L. (2011), “Dependent Hierarchical Beta Process for Image Interpolation and Denoising,” in *AISTATS*.